# NLP Scholar: A Dataset for Examining the State of NLP Research

**Saif M. Mohammad**
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

**Abstract**

Google Scholar is the largest web search engine for academic literature that also provides access to rich metadata associated with the papers. The ACL Anthology (AA) is the largest repository of articles on Natural Language Processing (NLP). We extracted information from AA for about 44 thousand NLP papers and identified authors who published at least three papers in AA. We then extracted citation information from Google Scholar for all their papers (not just their AA papers). This resulted in a dataset of 1.1 million papers and associated Google Scholar information. We aligned the information in the AA and Google Scholar datasets to create the *NLP Scholar Dataset*—a single unified source of information (from both AA and Google Scholar) for tens of thousands of NLP papers. NLP Scholar can be used to identify broad trends in productivity, focus, and impact of NLP research. We present here initial work on analyzing the volume of research in NLP over the years and identifying the most cited papers in NLP. We also list a number of additional potential applications.

**Keywords:** Scientometrics, Trends in Research, Google Scholar, ACL Anthology, Citations

## 1. Introduction

Google Scholar is a free web search engine for academic literature—peer reviewed journals, conferences, preprints, patents, theses, technical reports, etc.[1] Through it, users can access the metadata associated with an article and often the full text of the article as well. A key aspect of the metadata is the number of citations that an article has received.

Google Scholar does not provide information on how many articles are included in its database. However, scientometric researchers have estimated that it included about 389 million documents in January 2018 (Gusenbauer, 2019)—making it the world's largest source of academic information.[2] Thus, it is not surprising that there is growing interest in the use of Google Scholar information to draw inferences about scholarly research (Martín-Martín et al., 2018; Mingers and Leydesdorff, 2015; Orduña-Malea et al., 2014; Howland, 2010).

Our interests lie in the study of scholarly research in Natural Language Processing (NLP). However, Google Scholar does not provide information on the field of study pertaining to individual papers.[3] Thus, in this project, we combine information from Google Scholar with a dedicated high-quality source of information for NLP papers, the ACL Anthology (AA).

The ACL Anthology is a digital repository of public domain, free to access, articles on Natural Language Processing (NLP).[4] It includes papers published in the family of ACL conferences as well as in other NLP conferences such as LREC and RANLP.[5] When it was first launched in 2002, it provided access to rich metadata and full text of about 3,100 NLP papers. As of June 2019, it included close to 50,000 articles published since 1965—the year of the first ACL conference. It is the largest single source of scientific literature on NLP.

In this paper, we present the *NLP Scholar Dataset*—a single unified source of information (from both AA and Google Scholar) for tens of thousands of NLP papers. In Section 3, we present details on how the dataset was created and what it includes. Note that while AA is freely available, the information about papers is spread across several files. Through NLP Scholar, we not only aggregate the AA information into a standard database, we also add to the metadata, for example, by determining whether a paper is long paper or a short paper, whether it is a workshop paper or a main conference paper, whether it is a demo paper, etc. Further, even though Google Scholar information is freely available online, it is somewhat challenging to extract citation information for tens of thousands of papers. Thus we hope that NLP Scholar will save time and effort for other researchers. NLP Scholar is freely available from the project homepage.[6]

The NLP Scholar Dataset has numerous applications. Most notably, it can be used to examine the NLP literature to identify broad trends in productivity, focus, and impact of NLP research. We present here work on analyzing the volume of NLP research and on identifying some of the most cited papers in NLP in Sections 4 and 5, respectively. The analyses are presented as a sequence of questions and answers. Our broader goal here is simply to record the state of the NLP literature: who and how many of us are publishing? what are we publishing on? where and in what form are we publishing? and what is the impact of our publications? The answers are usually in the form of numbers, graphs, and visualizations. In Section 6, we list additional applications, before presenting concluding remarks in Section 7

Additional work that uses the NLP Scholar dataset for specific analyses is presented in separate papers. Moham-

[1] https://scholar.google.com

[2] Scientometrics is the study of scientific literature using quantitative techniques.

[3] Google Scholar does allow scholars to provide up to five tags for their profile corresponding to their area of research. However, the use of this feature by scholars is not consistent. Further, one scholar may publish papers on several fields of study.

[4] https://www.aclweb.org/anthology/

[5] ACL licenses its papers with a Creative Commons Attribution 4.0 International License.

[6] http://saifmohammad.com/WebPages/nlpscholar.html

mad (2020a) presents a comprehensive overview of citations in NLP Literature. Specifically, it explores questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers published in different time spans? how well cited are papers from different areas of research within NLP etc. Mohammad (2020b) quantifies and examines gender gap in Natural Language Processing Research; specifically, the disparities in authorship and citations across gender. Some of the analyses presented in the papers associated with this project are also available as a series of blog posts online.[7]

## 2. Related Work

This work is inspired by a vast amount of past research, including that on Google Scholar (Khabsa and Giles, 2014; Howland, 2010; Orduña-Malea et al., 2014; Martín-Martín et al., 2018), on the analysis of NLP papers (Radev et al., 2016; Anderson et al., 2012; Bird et al., 2008; Schluter, 2018; Mariani et al., 2018; Qazvinian et al., 2013; Teich, 2010; Saggion et al., 2017), on citation intent (Aya et al., 2005; Teufel et al., 2006; Pham and Hoffmann, 2003; Nanba et al., 2011; Mohammad et al., 2009; Zhu et al., 2015), and on measuring scholarly impact (Ravenscroft et al., 2017; Priem and Hemminger, 2010; Bulaitis, 2017; Bos and Nitza, 2019; Ioannidis et al., 2019; Yogatama et al., 2011; Mishra et al., 2018).

## 3. Data

We extracted information from both the ACL Anthology and Google Scholar in June 2019.[8] The three subsections below describe the information extracted from AA, the information extracted from Google Scholar, and how we aligned the information.

### 3.1. The ACL Anthology Data

AA provides access to its data through its website and a github repository.[9] The code for the ACL Anthology service is open source and available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License (Gildea et al., 2018).[10] The github repository includes a "data/xml/" directory that houses individual xml files for each of the proceedings. Table 1 shows the primary information we extracted from each of these xml files.

**Heuristics to obtain secondary information from AA:** AA does not explicitly record certain attributes of the paper such as whether it is a main conference paper, a student research paper, a system demonstration paper, a shared task paper, a workshop paper, a tutorial abstract, etc. It also does not record whether it is a long paper or a short paper. Since such attributes allow for interesting analyses of the data, we employ simple heuristics to determine their values

| Attribute | Example |
|---|---|
| paper-id | P18-1017 |
| paper-title | Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words |
| paper-url | https://www.aclweb.org/anthology/P18-1017 |
| paper-doi | 10.18653/v1/P18-1017 |
| volume-id | 1 |
| venue-code | P |
| booktitle | Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics |
| publisher | Association for Computational Linguistics |
| address | Melbourne, Australia |
| month | July |
| year | 2018 |
| author list | Saif M. Mohammad |

Table 1: Primary information extracted from AA. Here *year* stands for *year of publication*.

(secondary information) from the primary information already extracted from AA. Specifically, we determine values for attributes (such as those mentioned above) by searching for patterns in the booktitle—see Table 2. The attribute has value 1 if the pattern is found, and 0 otherwise.

**Authors:** The xml files include author names only from the year 2013 and later. This is likely when ACL conferences began to explicitly ask for author names (through a text box) at the submission page. However, AA also provides a master BibTeX file that includes the BibTeX entries for each of the papers in AA. Much of the information in the BibTeX entries is already present in the xml files, but notably it includes the author list for all papers (and not just those published since 2013). Thus we extract author names from the master BibTeX file.

Multiple authors can have the same name and the same authors may use multiple variants of their names in papers. This presents a problem that all paper aggregation projects have to face, including AA and Google Scholar. The AA volunteer team handles such ambiguities using both semi-automatic and manual approaches (fixing some instances on a case by case basis). In cases of multiple authors with the same name (which occurs very infrequently in AA) the authors are manually assigned an author id. Additionally, AA keeps a file that includes canonical forms of author names as well as name variants. We use AA's information on authors to transform author names obtained from the BibTeX entries into canonical forms.

**Venue Code:** AA assigns a unique venue code to each of the conferences and journals (e.g., the venue code for ACL is P), whereas all workshops get the venue code W.[11] We use a mapping file included in the AA repository to map the venue codes to the venue names (e.g., P to ACL).[12]

**Number of papers in AA:** As of June 2019, AA had ∼50K entries, however, this includes some number of entries that

[7] https://medium.com/@nlpscholar/state-of-nlp-cbf768492f90

[8] Thus, all subsequent papers and citations are not included in the analysis. A fresh data collection is planned for January 2020.

[9] https://www.aclweb.org/anthology/ https://github.com/acl-org/acl-anthology

[10] https://creativecommons.org/licenses/by-nc-sa/3.0/

[11] The venue code system is scheduled to be updated in 2020 when the letter system is to be replaced by venue abbreviations.

[12] Note that the distinction between a conference and a workshop can sometimes be fuzzy. Further, some venues (e.g. EMNLP) started off as workshops, but eventually gained a conference status. For this work, we treated them as conferences (even for the earlier years).

| Attribute | Pattern | Example |
|---|---|---|
| demo paper | *demo* | 0 |
| main conference paper | *conference* | 1 |
| shared task paper | *shared task* | 0 |
| short paper | *short papers* | 0 |
| student workshop paper | *student* | 0 |
| tutorial flag | *tutorial* | 0 |
| workshop paper | *workshop* | 0 |

Table 2: Secondary information obtained using simple heuristics. The attribute has value 1 if the pattern is found in the book title, and 0 otherwise. Example data is for the same paper as in Table 1.

are not truly research publications (for example, forewords, prefaces, table of contents, programs, schedules, indexes, calls for papers/participation, lists of reviewers, lists of tutorial abstracts, invited talks, appendices, session information, obituaries, book reviews, newsletters, lists of proceedings, lifetime achievement awards, erratum, and notes). We discard them for the analyses here. (Note: CL journal includes position papers, letters to editor, opinions, etc. We do not discard them.) We are then left with 44,896 articles.

### 3.2. Google Scholar Data

Google Scholar was launched in November 2004 and has undergone several rounds of refinements since. Notably, since 2012, it allowed scholars/researchers to create and edit public author profiles *Scholar Citations Profiles*. GS then presented the number of citations to their articles on a profile page, and also calculated metrics such as total number of citations, h-index, and i-10 index.

GS does not provide an API to extract information about the papers. This is likely because of its agreement with publishing companies that have scientific literature behind paywalls (Martín-Martín et al., 2018). We extracted citation information from Google Scholar profiles of authors who published in the ACL Anthology. This is explicitly allowed by GS's robots exclusion standard. This is also how past work has studied Google Scholar (Khabsa and Giles, 2014; Orduña-Malea et al., 2014; Martín-Martín et al., 2018).

We extracted citation information from Google Scholar profiles of authors who had a Google Scholar Profile page and had published at least three papers in the ACL Anthology. This yielded citation information for 1.1 million papers in total. We will refer to this dataset as the *NLP Subset of the Google Scholar Dataset*, or *GScholar-NLP* for short. Note that GScholar-NLP includes citation information not just for NLP papers, but also for non-NLP papers published by authors who have at least three papers in AA. Table 3 shows the information that was extracted. Here *year* stands for *year of publication*. *pubid* is an internal id used by Google Scholar.

GScholar-NLP includes 33,051 of the 44,896 papers in AA (about 75%). We will refer to this subset of the ACL Anthology papers as AA'. The citation analysis presented later in this paper are on AA'.

| Attribute | Example |
|---|---|
| title | Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words |
| authors | S Mohammad |
| conference | Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics |
| cites | 24 |
| year | 2018 |
| pubid | h168fVGZblEC |

Table 3: Primary information extracted from Google Scholar.

### 3.3. Aligning Data from the ACL Anthology and Google Scholar

The ACL Anthology and Google scholar do not share a common paper id system. Thus we needed a shared piece of information, that is unique to the paper, to create a single unified dataset (with information from both AA and GS).

Often the paper title is unique and can be used as a paper id. However, there are a small number of instances where two (or more) different papers have the same title. Thus, for this work, we use a concatenation of the paper title and year of publication as the paper id. Further, we remove all non-alphanumeric characters from this paper id. This is because occasionally there might be differences in how the title is stored for the same paper in AA and Google Scholar because of hyphens, spaces, and special characters. Thus, for example, the paper id for the example paper listed in Tables 1 through 3 would be: *obtainingreliablehumanratingsofvalencearousalanddominancefor20000englishwords2018*. We use this title–year concatenation as the paper id to align information between AA and GS, and also to align information spread across different AA sources such as the xml files and the master BibTeX file.

We avoid using author names in the paper id because Google only stores the last name and initials of the first and middle names (as shown in Table 3). Further, special characters are much more common in author names (than in titles or years) and thus it is much more likely that the same author is written in a different form across AA and Google Scholar.

However, in case the paper is submitted to a preprint server (such as ArXiv) in a different year than the year of publication at an AA venue, there can be a mismatch between the years of publications as recorded in AA and GS. In such cases, we use the paper id formed by the concatenation of the title and the last name of the first author.

We refer to the final combined information from AA and GS as the *NLP Scholar Dataset*. All heuristics described above were chosen for high precision, which was sometimes at the cost of recall. Several rounds of spot checks by the author were followed by the distribution of the data and visualizations to outside researchers for feedback. Nonetheless, it should be noted that NLP Scholar inherits errors from AA and Google Scholar, and might still have a small number of alignment errors caused by boundary cases. A detailed compilation of the caveats and limitations is available online in the *About NLP Scholar* page.[13]

[13] https://medium.com/@nlpscholar/about-nlp-scholar-62cb3b0f4488

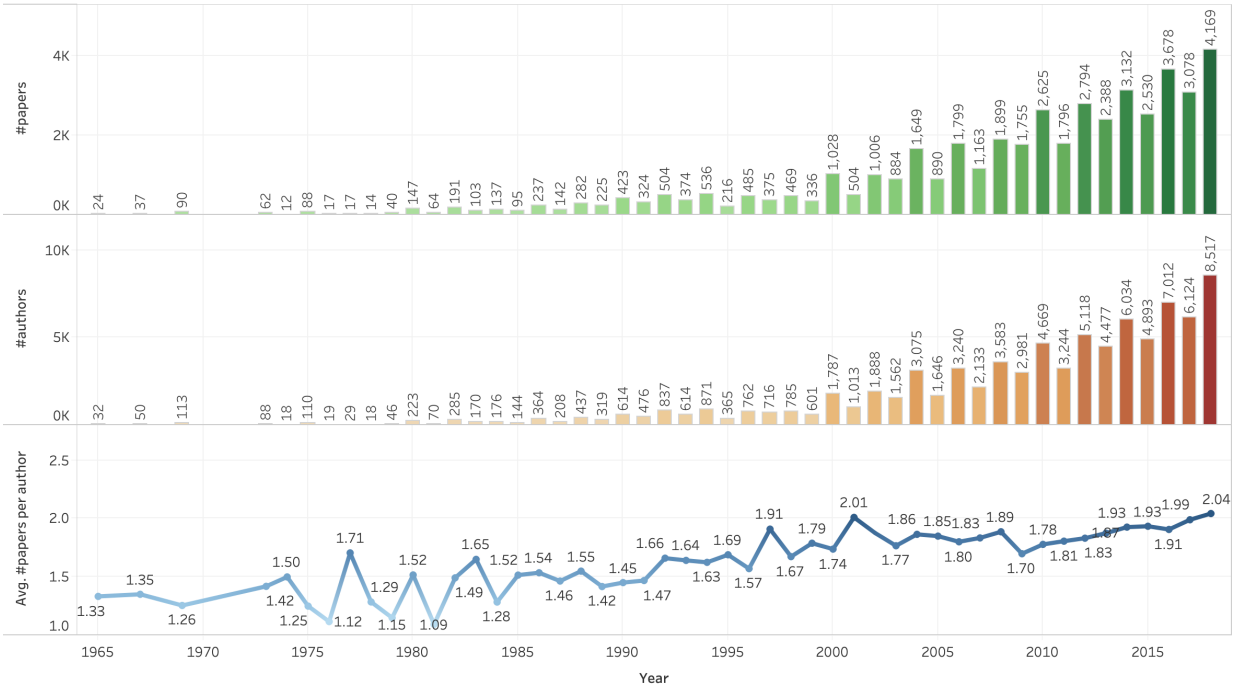


Figure 1: ACL Anthology (1965 to 2018): number of papers, number of authors, and average number of papers per author.

## 4. The Volume of NLP Research

The volume of research in a field is a simple yet powerful indicator of the health of the field. It provides a window into questions such as: Is the field energized and drawing new researchers at greater numbers than before?; How productive are the researchers in the field?; What percentage of the researchers leave after publishing just one paper?; What types of papers are published more in the field?; etc. We explore these questions for Natural Language Processing using the NLP Scholar Dataset. It should be noted that there exist NLP papers outside of AA. However, since AA is the single largest source of NLP papers, it is likely that the analyses below shed light not just on AA papers but also, to some extent, on NLP research in general.

*Q1. How big is the ACL Anthology (AA)? How is the number of papers changing with time?*

A. As mentioned in the Data section, as of June 2019, AA had ∼50K entries, however, after excluding non-paper entries, we are left with 44,896 articles. The top graph in Figure 1 shows the number of papers published in each of the years from 1965 to 2018.

Discussion: Observe that there was a spurt of papers in the 1990s, but a much larger growth has occurred since the year 2000. Also, note that the number of publications is considerably higher in alternate years. This is due to biennial conferences. Since 1998 the largest of such conferences has been LREC. (In 2018 alone LREC had over 700 main conferences papers and additional papers from its 29 workshops). COLING, another biennial conference (also occurring in the even years) has about 45% of the number of main conference papers as LREC.

*Q2. How many authors/researchers publish in the ACL Anthology?*

A. The 44,896 articles in AA have 37,300 authors. Figure 1 (middle graph) shows the number of authors by year:

Discussion: Observe that the number of authors is also growing over time (mirroring the number of papers). Continually attracting a greater number of researchers is a sign of good health for natural language research.

*Q3. Are we publishing more papers per author now than in earlier decades?*

A. Yes. The average number of papers per author is the highest it has ever been in 2018. Although, there have been other years where the average has been close to the highest (as in 2001). The bottom graph in Figure 1 shows the average number of papers per author from 1965 to 2018.

Discussion: One can observe the general trend of increasing number of papers per author from 2009 to 2018. One can also observe that the period between 1965 and 1991 had a markedly lower average compared to the period between 1992 and 2018, despite some outliers. We see large fluctuations in the numbers for the 1965–1990 period; this is likely because of the small number of papers in those years.

*Q4. How many authors published exactly one paper in AA?*

A. 21,606 authors have published exactly one paper in AA. This is about 58% of the total number of authors who have published in AA. Consider these bins of papers in AA: 1 (the author has published exactly one paper in AA); 2 to 10 (the author has published two to ten papers in AA); 10 to 100 (the author has published ten to one hundred papers in AA); 100+ (the author has published more than one

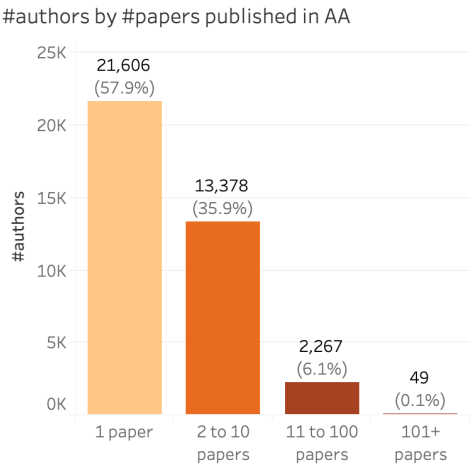


Figure 2: Number of authors by number of papers.

hundred papers in AA). Figure 2 shows the number of AA authors in each bin.

Discussion: It is interesting to note that a majority of authors have exactly one paper in AA. While we do not have information on how many of these authors have NLP papers outside of AA, it is still likely that a large portion of those that publish NLP papers only publish one NLP paper. Further work is needed to determine if this is common in other anthologies as well. Further work is also needed to determine whether this is a healthy percentage—in terms of the success of attracting new people (especially students) to the field vs. the lack of success in enabling more people to publish beyond their first NLP paper.

*Q5. How many people are actively publishing in NLP?*

A. It is hard to know the exact number, but we can determine the number of people who have published in AA in the last N years.

> #people who published at least one paper in 2017 and 2018 (2 years): 11,957
> #people who published at least one paper 2015 through 2018 (4 years):17,457

Of course, some number of researchers published NLP papers in non-AA venues.

*Q6. How many journal papers exist in the AA? How many main conference papers? How many workshop papers?*

A. See Figure 3.

Discussion: The number of journal papers is dwarfed by the number of conference and workshop papers. (This is common in computer science. Even though NLP is a broad interdisciplinary field, the influence of computer science practices on NLP is particularly strong.) Shared task and system demo papers are relatively new (introduced in the 2000s), but their numbers are already substantial.

Creating a separate class for "Top-tier Conference" is somewhat arbitrary, but it helps make certain comparisons more meaningful (for example, when comparing the

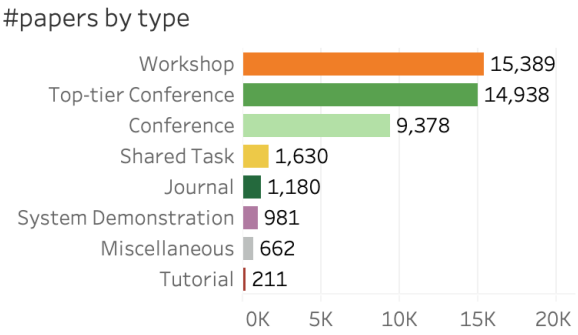


Figure 3: Number of AA papers by type.

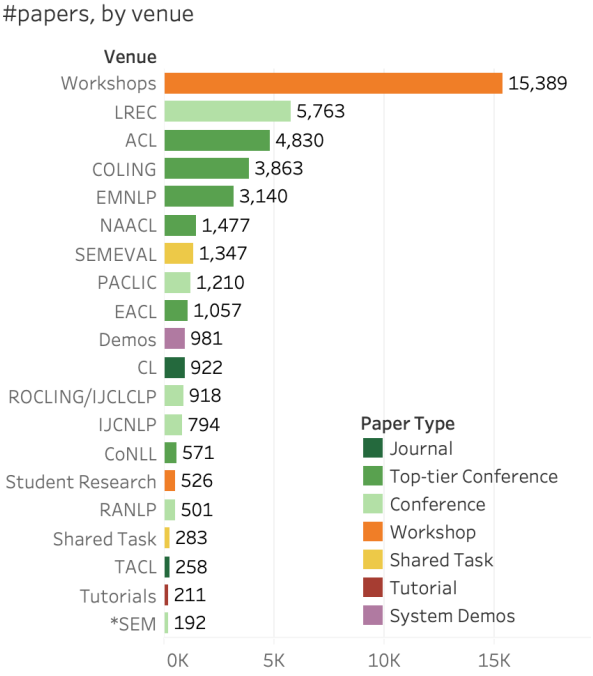


Figure 4: The number of main conference papers for various venues and paper types (workshop papers, demos, etc.).

average number of citations, etc.). For this work, we consider ACL, EMNLP, NAACL, COLING, and EACL as top-tier conferences based on low acceptance rates and high citation metrics, but certainly other groupings are also reasonable.

*Q7. How many papers have been published at ACL (main conference papers)? What is the distribution of the number of papers across various NLP venues?*

A. # ACL (main conference papers) as of June 2018: 4,830

The same workshop can co-occur with different conferences in different years, so we grouped all workshop papers in their own class. We did the same for tutorials, system demonstration papers (demos), and student research papers. Figure 4 shows the number of main conference papers for various venues and paper types (workshop papers, demos, etc.).

Discussion: Even though LREC is a relatively new conference that occurs only once in two years, it tends to have a high acceptance rate ($\sim$60%), and enjoys substantial participation. Thus, LREC is already the largest single source of NLP conference papers. SemEval, which started as SenseEval in 1998 and occurred once in two or three years, has now morphed into an annual two-day workshop—SemEval. It is the largest single source of NLP shared task papers.

## 5. Most Cited Papers

Research articles can have impact in a number of ways—pushing the state of the art, answering crucial questions, finding practical solutions that directly help people, making a new generation of potential-scientists excited about a field of study, and more. However, measures of research impact are limited in scope; they measure only some kinds of contributions.

The most commonly used metrics of research impact are derived from citations. A citation of a scholarly article is the explicit reference to that article. Several citation-based metrics have emerged over the years including: number of citations, average citations, h-index, relative citation ratio, and impact factor.

It is not always clear why some papers get lots of citations and others do not. One can argue that highly cited papers have captured the imagination of the field: perhaps because they were particularly creative, opened up a new area of research, pushed the state of the art by a substantial degree, tested compelling hypotheses, or produced useful datasets, among other things. Note however, that the number of citations is not always a reflection of the quality or importance of a piece of work. Note also that there are systematic biases that prevent certain kinds of papers from accruing citations, especially when the contributions of a piece of work are atypical, not easily quantified, or in an area where the number of scientific publications is low. Further, the citations process can be abused, for example, by egregious self-citations (Ioannidis et al., 2019).

Nonetheless, given the immense volume of scientific literature, the relative ease with which one can track citations using services such as Google Scholar and Semantic Scholar, and given the lack of other easily applicable and effective metrics, citation analysis is an imperfect but useful window into research impact.

In this section, we examine the most cited papers in AA'. As mentioned earlier, AA' is a subset of AA, for which we were able to extract citation information from Google Scholar. It includes 33,051 out of the 44,896 papers in AA. The papers in AA' received $\sim$1.2 million citations (as of June 2019). Figure 5 shows a timeline graph where each year has a bar with height corresponding to the number of citations received by papers published in that year. Further, the bar has colored fragments corresponding to each of the papers and the height of a fragment (paper) is proportional to the number of citations it has received. Thus it is easy to spot the papers that received a large number of citations, and the years when the published papers received a large number of citations.

Figure 5 is a screenshot of an interactive visualization of the data. Hovering over individual papers reveals an information box showing the paper title, authors, year of publication, publication venue, and #citations. The interactive visualization is freely available from the project homepage.[14]

Observe that with time, not only have the number of papers grown, but also the number of high-citation papers. We see a marked jump in the 1990s over the previous decades, but the 2000s are the most notable in terms of the high number of citations. The 2010s papers will likely surpass the 2000s papers in the years to come.

Through the questions below we explore the most cited papers in AA': overall and across specific subsets of AA'.

*Q1. What are the most cited papers in AA'?*

A. Figure 6 shows the most cited papers in the AA'.

Discussion: We see that the top-tier conference papers (green) are some of the most cited papers in AA'. There are a notable number of journal papers (dark green) in the most cited list as well, but very few demo (purple) and workshop (orange) papers.

In the interactive visualizations (to be released later), one can click on the url to be taken directly to the paper's landing page in the ACL Anthology website. That page includes links to metadata, the pdf, and associated files such as videos and appendices. There will also be functionality to download the lists.

*Q2. What are the most cited AA' journal papers? What are the most cited AA' workshop papers? What are the most cited AA' shared task papers? What are the most cited AA' demo papers? What are the most cited tutorials? What are the most cited papers from individual venues such as ACL, LREC, and EMNLP?*

A. Figure 7 shows the most cited journal papers in the AA'. Individual lists of the most cited AA' conference papers, workshop papers, system demo papers, shared task papers, and tutorials can be viewed online.[15] (These lists are not shown here due to space constraints.) The most cited papers from individual venues (ACL, CL journal, TACL, EMNLP, LREC, etc.) can also be viewed there.

Discussion: Machine translation papers are well-represented in many of these lists, but especially in the system demo papers list. Toolkits such as MT evaluation ones, NLTK, Stanford Core NLP, WordNet Similarity, and OpenNMT have highly cited demo or workshop papers.

The shared task papers list is dominated by task description papers (papers by task organizers describing the data and task), especially for sentiment analysis tasks. However, the list also includes papers by top-performing systems in these shared tasks, such as the NRC-Canada, HidelTime, and UKP papers.

[14] http://saifmohammad.com/WebPages/nlpscholar.html

[15] https://medium.com/@nlpscholar/the-state-of-nlp-literature-part-iiia-845eb5dc3364

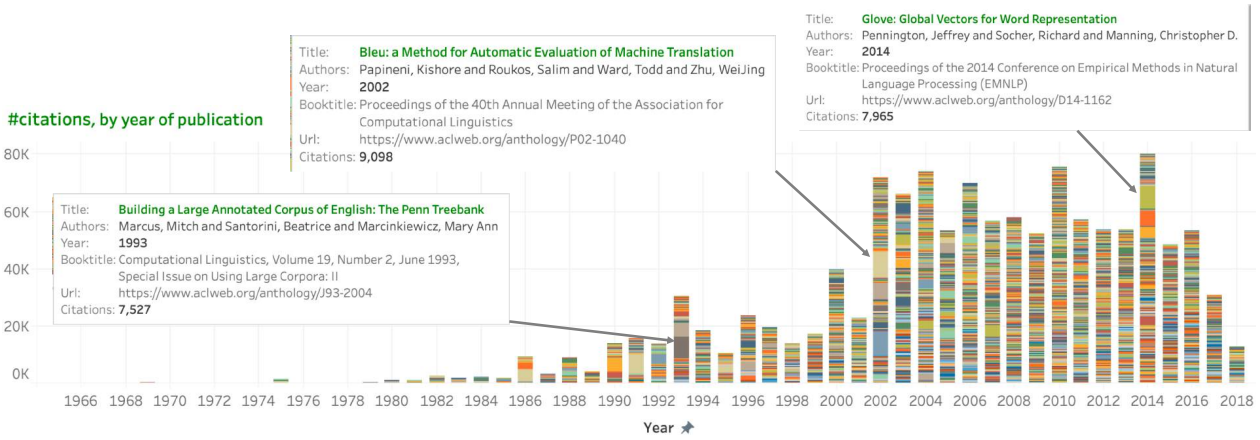


Figure 5: A timeline graph where each year has a bar with height corresponding to the number of citations received by papers published in that year. The bar has colored fragments corresponding to each of the papers and the height of a fragment (paper) is proportional to the number of citations it has received.

| Paper-Id | Paper-Title | Author(s) | Year | Url (click for pdf) | |
|---|---|---|---|---|---|
| P02-1040 | Bleu: a Method for Automatic Evaluation of Machine .. | Papineni, Kishore and Roukos, Salim a.. | 2002 | https://www.ac.. | 9,098 |
| W02-1011 | Thumbs up? Sentiment Classification using Machine L.. | Pang, Bo and Lee, Lillian and Vaithyan.. | 2002 | https://www.ac.. | 8,187 |
| D14-1162 | Glove: Global Vectors for Word Representation | Pennington, Jeffrey and Socher, Rich.. | 2014 | https://www.ac.. | 7,965 |
| J93-2004 | Building a Large Annotated Corpus of English: The Pe.. | Marcus, Mitch and Santorini, Beatrice.. | 1993 | https://www.ac.. | 7,527 |
| J91-4003 | The Generative Lexicon | Pustejovsky, James | 1991 | https://www.ac.. | 6,593 |
| P02-1053 | Thumbs Up or Thumbs Down? Semantic Orientation A.. | Turney, Peter | 2002 | https://www.ac.. | 5,642 |
| D14-1179 | Learning Phrase Representations using RNN Encoder.. | Cho, Kyunghyun and van Merrienboer.. | 2014 | https://www.ac.. | 5,344 |
| J93-2003 | The Mathematics of Statistical Machine Translation: .. | Brown, Peter F. and Della Pietra, Step.. | 1993 | https://www.ac.. | 5,047 |
| J90-1003 | Word Association Norms, Mutual Information, and Le.. | Church, Kenneth and Hanks, Patrick | 1990 | https://www.ac.. | 4,845 |
| P07-2045 | Moses: Open Source Toolkit for Statistical Machine Tr.. | Koehn, Philipp and Hoang, Hieu and Bi.. | 2007 | https://www.ac.. | 4,581 |
| D14-1181 | Convolutional Neural Networks for Sentence Classific.. | Kim, Yoon | 2014 | https://www.ac.. | 4,362 |
| J86-3001 | Attention, Intentions, and the Structure of Discourse | Grosz, Barbara J. and Sidner, Candace.. | 1986 | https://www.ac.. | 4,101 |
| J03-1002 | A Systematic Comparison of Various Statistical Align.. | Och, Franz Josef and Ney, Hermann | 2003 | https://www.ac.. | 4,040 |
| C92-2082 | Automatic Acquisition of Hyponyms from Large Text .. | Hearst, Marti A. | 1992 | https://www.ac.. | 3,749 |
| P14-5010 | The Stanford CoreNLP Natural Language Processing T.. | Manning, Christopher D. and Surdean.. | 2014 | https://www.ac.. | 3,543 |

Figure 6: The fifteen most cited papers in AA'.

| Paper-Id | Paper-Title | Author(s) | Year | Url (click for pdf) | |
|---|---|---|---|---|---|
| J93-2004 | Building a Large Annotated Corpus of English: The Pe.. | Marcus, Mitch and Santorini, Beatrice.. | 1993 | https://www.ac.. | 7,527 |
| J91-4003 | The Generative Lexicon | Pustejovsky, James | 1991 | https://www.ac.. | 6,593 |
| J93-2003 | The Mathematics of Statistical Machine Translation: .. | Brown, Peter F. and Della Pietra, Step.. | 1993 | https://www.ac.. | 5,047 |
| J90-1003 | Word Association Norms, Mutual Information, and Le.. | Church, Kenneth and Hanks, Patrick | 1990 | https://www.ac.. | 4,845 |
| J86-3001 | Attention, Intentions, and the Structure of Discourse | Grosz, Barbara J. and Sidner, Candace.. | 1986 | https://www.ac.. | 4,101 |
| J03-1002 | A Systematic Comparison of Various Statistical Align.. | Och, Franz Josef and Ney, Hermann | 2003 | https://www.ac.. | 4,040 |
| J96-2004 | Assessing Agreement on Classification Tasks: The Ka.. | Carletta, Jean | 1996 | https://www.ac.. | 2,429 |
| J03-4003 | Head-Driven Statistical Models for Natural Language .. | Collins, Michael | 2003 | https://www.ac.. | 2,271 |
| J05-1004 | The Proposition Bank: An Annotated Corpus of Seman.. | Palmer, Martha and Gildea., Daniel an.. | 2005 | https://www.ac.. | 2,164 |
| J90-2002 | A Statistical Approach to Machine Translation | Brown, Peter F. and Cocke, John and .. | 1990 | https://www.ac.. | 2,102 |
| J11-2001 | Lexicon-Based Methods for Sentiment Analysis | Taboada, Maite and Brooke, Julian an.. | 2011 | https://www.ac.. | 1,982 |
| J02-3001 | Automatic Labeling of Semantic Roles | Gildea., Daniel and Jurafsky, Dan | 2002 | https://www.ac.. | 1,956 |
| J06-1003 | Evaluating WordNet-based Measures of Lexical Sema.. | Budanitsky, Alexander and Hirst, Gra.. | 2006 | https://www.ac.. | 1,750 |
| J93-1004 | A Program for Aligning Sentences in Bilingual Corpora | Gale, William A. and Church, Kenneth | 1993 | https://www.ac.. | 1,573 |
| Q17-1010 | Enriching Word Vectors with Subword Information | Bojanowski, Piotr and Grave, Édouard.. | 2017 | https://www.ac.. | 1,516 |

Figure 7: The fifteen most cited journal papers in AA'.

| Paper-Id | Paper-Title | Author(s) | Year | Url (click for pdf) | |
|---|---|---|---|---|---|
| D14-1162 | Glove: Global Vectors for Word Representation | Pennington, Jeffrey and Socher, Rich.. | 2014 | https://www.ac.. | 7,965 |
| D14-1179 | Learning Phrase Representations using RNN Encoder.. | Cho, Kyunghyun and van Merrienboer.. | 2014 | https://www.ac.. | 5,344 |
| D14-1181 | Convolutional Neural Networks for Sentence Classific.. | Kim, Yoon | 2014 | https://www.ac.. | 4,362 |
| P14-5010 | The Stanford CoreNLP Natural Language Processing T.. | Manning, Christopher D. and Surdean.. | 2014 | https://www.ac.. | 3,543 |
| D13-1170 | Recursive Deep Models for Semantic Compositionalit.. | Socher, Richard and Perelygin, Alex a.. | 2013 | https://www.ac.. | 2,798 |
| L10-1-531 | SentiWordNet 3.0: An Enhanced Lexical Resource for .. | Baccianella, Stefano and Esuli, Andre.. | 2010 | http://www.lre.. | 2,263 |
| N13-1090 | Linguistic Regularities in Continuous Space Word Rep.. | Mikolov, Tomáš and Yih, Wentau and .. | 2013 | https://www.ac.. | 2,081 |
| J11-2001 | Lexicon-Based Methods for Sentiment Analysis | Taboada, Maite and Brooke, Julian an.. | 2011 | https://www.ac.. | 1,982 |
| D15-1166 | Effective Approaches to Attention-based Neural Mac.. | Luong, Minh-Thang and Pham, Hieu a.. | 2015 | https://www.ac.. | 1,961 |
| P14-1062 | A Convolutional Neural Network for Modelling Sente.. | Kalchbrenner, Nal and Grefenstette, .. | 2014 | https://www.ac.. | 1,794 |
| P10-1040 | Word Representations: A Simple and General Method.. | Turian, Joseph and Ratinov, LevArie a.. | 2010 | https://www.ac.. | 1,753 |
| W14-4012 | On the Properties of Neural Machine Translation: Enc.. | Cho, Kyunghyun and van Merrienboer.. | 2014 | https://www.ac.. | 1,673 |
| Q17-1010 | Enriching Word Vectors with Subword Information | Bojanowski, Piotr and Grave, Édouard.. | 2017 | https://www.ac.. | 1,516 |
| N16-3020 | "Why Should I Trust You?": Explaining the Predictions.. | Ribeiro, Marco Tulio and Singh, Same.. | 2016 | https://www.ac.. | 1,387 |
| W11-0705 | Sentiment Analysis of Twitter Data | Agarwal, Apoorv and Xie, Boyi and Vo.. | 2011 | https://www.ac.. | 1,369 |

Figure 8: The fifteen most cited AA' papers from the 2010s.

*Q3. What are the most cited AA' papers published in the last decade? What are the most cited papers in various time periods from the past?*

A. Figure 8 shows the most cited AA' papers from the last decade. The most cited AA' papers published from various other time spans are available online.[16]

Discussion: The early period (1965–1989) includes papers on grammar and linguistic structure. The 1990s list has papers addressing many different NLP problems with statistical approaches. Papers on MT and sentiment analysis are frequent in the 2000s list. The 2010s are dominated by papers on word embeddings and neural representations.

## 6. Further Explorations with NLP Scholar

The NLP Scholar Dataset has several uses, including but not limited to those listed below. Some of these uses involve assisting studies in better understanding the NLP research landscape. Other uses are for developing practical applications.

- NLP is a diverse inter-disciplinary field where one's research may be contributing to (and may have been influenced by) various other fields such as psychology, humanities, and social sciences. Thus an interesting question is—what makes a paper an NLP paper? The NLP Scholar dataset and the larger GScholar-NLP dataset (papers by NLP authors) can be used to explore this question.
- What makes papers highly cited? We have historical citation information. Can we use that to identify notable characteristics of high-citation papers?
- Tracking disparities in the number of authors from various demographic groups. For example, tracking the participation of women in NLP research.
- Identifying better matches of reviewers with papers and mentors with student researchers.
- Tracking disparities in citations across various demographic groups. For example, determining and tracking whether women are cited more or less than men in NLP research.
- Determining the average citations impact of different types of papers and venues: e.g. how influential are system demo papers?; how well cited are workshop papers compared to main conference papers?; etc.
- Identifying related work. Given a query term, provide papers that are relevant.
- Assigning topics to papers and tracking topics over time. Again this can help with the related work search, but also with tracking popularity of topics over time.
- Quantifying the impact of non-NLP fields on NLP. One way is to identify how often we cite papers from non-NLP fields. Which fields are we citing a lot? How is that changing with time? Another avenue is to examine the GScholar-NLP dataset and identify how often the NLP authors publish papers in non-NLP fields.

The NLP Scholar Dataset and associated interactive visualizations are freely available from the project homepage.[17]

## 7. Conclusions

We aligned the information in the ACL Anthology and Google Scholar to create the *NLP Scholar Dataset*—a single unified source of information (from both AA and Google Scholar) for tens of thousands of NLP papers. NLP Scholar can be used to examine the literature as a whole to identify broad trends in productivity, focus, and impact. We presented initial work on analyzing the volume of NLP research and the most cited papers in NLP. We showed that not only are the number of papers and authors publishing in AA growing over time, the number of papers published by an author in a year is also steadily increasing. We also showed that a majority of authors publish just one paper in AA. We created various lists of most cited AA papers (overall and by type) and qualitatively discussed the trends in these lists. Finally, we listed a number of potential applications of the NLP Scholar and the GScholar-NLP datasets. The latter includes Google Scholar information for 1.1 million papers by authors who published at least three papers in AA. We hope that these resources will foster further research into various aspects of NLP research.

[16] https://medium.com/@nlpscholar/the-state-of-nlp-literature-part-iiia-845eb5dc3364

[17] http://saifmohammad.com/WebPages/nlpscholar.html

## Acknowledgments

This work was possible due to the helpful discussion and encouragement from a number of awesome people, including: Dan Jurafsky, Tara Small, Michael Strube, Cyril Goutte, Eric Joanis, Matt Post, Patrick Littell, Torsten Zesch, Ellen Riloff, Norm Vinson, Iryna Gurevych, Rebecca Knowles, Isar Nejadgholi, and Peter Turney. Also, a big thanks to the ACL Anthology Team for creating and maintaining a wonderful resource.

## 8. Bibliographical References

Anderson, A., McFarland, D., and Jurafsky, D. (2012). Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21. Association for Computational Linguistics.

Aya, S., Lagoze, C., and Joachims, T. (2005). Citation classification and its applications. In *Knowledge Management: Nurturing Culture, Innovation, and Technology*, pages 287–298. World Scientific.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., Lee, D., Powley, B., Radev, D. R., and Tan, Y. F. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

Bos, A. R. and Nitza, S. (2019). Interdisciplinary comparison of scientific impact of publications using the citation-ratio. *Data Science Journal*, 18(1).

Bulaitis, Z. (2017). Measuring impact in the humanities: Learning from accountability and economics in a contemporary history of cultural value. *Palgrave Communications*, 3(1):7.

Gildea, D., Kan, M.-Y., Madnani, N., Teichmann, C., and Villalba, M. (2018). The ACL anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia, July. Association for Computational Linguistics.

Gusenbauer, M. (2019). Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214.

Howland, J. L. (2010). How scholarly is google scholar? a comparison to library databases.

Ioannidis, J. P., Baas, J., Klavans, R., and Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS biology*, 17(8):e3000384.

Khabsa, M. and Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS one*, 9(5):e93949.

Mariani, J., Francopoulo, G., and Paroubek, P. (2018). The nlp4nlp corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3:36.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., and López-Cózar, E. D. (2018). Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4):1160–1177.

Mingers, J. and Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19.

Mishra, S., Fegley, B. D., Diesner, J., and Torvik, V. I. (2018). Self-citation is the hallmark of productive authors, of any gender. *PloS one*, 13(9):e0195773.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., and Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 584–592.

Mohammad, S. M. (2020a). Examining citations of natural language processing literature. In *Proceedings of the 2020 annual conference of the association for computational linguistics*, Seattle, USA.

Mohammad, S. M. (2020b). Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 2020 annual conference of the association for computational linguistics*, Seattle, USA.

Nanba, H., Kando, N., and Okumura, M. (2011). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1):117–134.

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., and López-Cózar, E. D. (2014). About the size of google scholar: playing the numbers. *arXiv preprint arXiv:1407.6239*.

Pham, S. B. and Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*, pages 759–771. Springer.

Priem, J. and Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social web. *First monday*, 15(7).

Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., and Moon, T. (2013). Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*, 46:165–201.

Radev, D. R., Joseph, M. T., Gibson, B., and Muthukrishnan, P. (2016). A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, 67(3):683–706.

Ravenscroft, J., Liakata, M., Clare, A., and Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PloS one*, 12(3):e0173152.

Saggion, H., Ronzano, F., Accuosto, P., and Ferrés, D. (2017). Multiscien: a bi-lingual natural language processing system for mining and enrichment of scientific collections. In *Mayr P, Chandrasekaran MK, Jaidka K, editors. Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natu-*

*ral Language Processing for Digital Libraries (BIRNDL 2017); 2017 Aug 11; Tokyo, Japan.[place unknown]: CEUR Workshop Proceedings; 2017. p. 26-40.* CEUR Workshop Proceedings.

Schluter, N. (2018). The glass ceiling in nlp. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798.

Teich, E. (2010). Exploring a corpus of scientific texts using data mining. In *Corpus-linguistic applications*, pages 233–247. Brill Rodopi.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110. Association for Computational Linguistics.

Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B. R., and Smith, N. A. (2011). Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 594–604. Association for Computational Linguistics.

Zhu, X., Turney, P., Lemire, D., and Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.